



Российская академия наук
Вычислительный центр имени А. А. Дородницына

Об алгоритмах прогнозирования процессов с плавно меняющимися закономерностями

Филипенков Николай Владимирович, к.ф.-м.н.,

Руководитель направления риск-менеджмента, SAS Россия / СНГ

Москва, 19 апреля 2011 г.



Краткая биография

Образование

- МГУ, Вычислительная Математика и Кибернетика
- Вычислительный центр РАН
к.ф.-м.н., 05.13.17 – «Теоретические основы информатики»

Опыт работы

- ОАО «Банк Москвы», 5 лет
Руководитель отдела анализа скорингового кредитования
- SAS
Руководитель направления риск-менеджмента

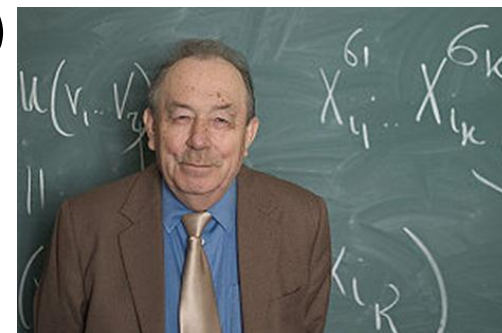
Интеллектуальный анализ данных

Data Mining — это процесс обнаружения в сырых данных следующих знаний:

- ранее неизвестных,
- нетривиальных,
- практически полезных,
- доступных интерпретации,
- необходимых для принятия решений в различных сферах человеческой деятельности.

Школа Ю.И. Журавлева

- Распознавание образов (pattern recognition)
 - Интеллектуальный анализ данных (data mining)
 - Машинное обучение (machine learning)
 - Дискретная математика
-
- Вычислительный центр РАН
 - Кафедра Математических методов прогнозирования ВМиК МГУ (с 1997)
 - Кафедра «Интеллектуальные системы» ФУПМ МФТИ (с 2004)



Конференции

- Математические методы распознавания образов
 - www.mmro.ru
- Интеллектуализация обработки информации
 - iip.mmro.ru
- Распознавание образов и анализ изображений
 - www.eltech.ru/pria2010/
- ГрафиКон
 - gc2010.graphicon.ru
- SAS A, SAS M
 - www.sas.com

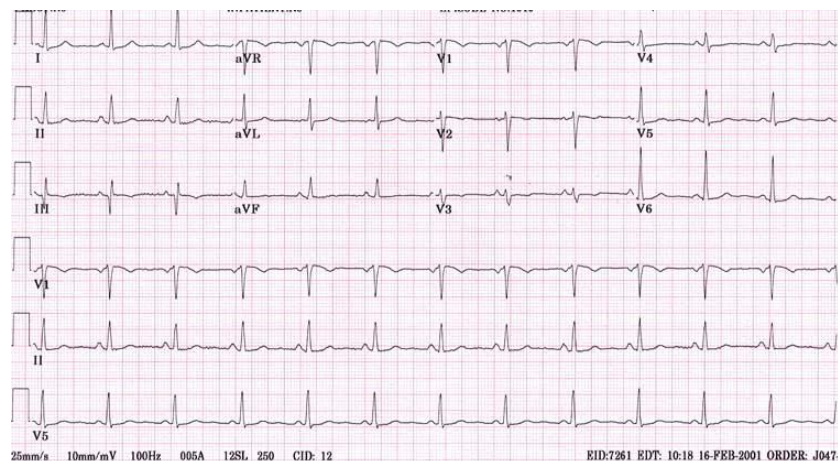
Анализ временных рядов

Области знаний

- Медицина
- Экономика
- Физика
- Химия
- ...

Временные ряды

- Многомерные
- Нестационарные
- ...



Методы анализа временных рядов

- Сглаживание и фильтрация (Р.Г. Браун)
- ARIMA (Дж. Бокс, Г. Дженкинс)
- GARCH (Р. Энгл)
- Спектральные методы (Д. Ваттс, Г. Дженкинс, Л. Заде, Дж. Рагаззини, Ф.Ф. Дедус)
- Статистические модели (С.А. Айвазян, В.М. Бухштабер, Т. Андерсон, М. Кендэл)
- Алгоритмы поиска правил (Р. Агравал, Г. Дас)
- Алгебраический подход к выделению трендов (Ю.И.Журавлёв, К.В. Рудаков, Ю.В. Чехович)

Цель работы

Разработать алгоритм поиска закономерностей в пучках временных рядов, базирующийся на предположении о плавном изменении закономерностей с течением времени

Постановка задачи

Пучок временных рядов

Время \longrightarrow		1	2	...	T-3	T-2	T-1	T
Ряд \downarrow	1	$a_{1,1}$	$a_{1,2}$...	$a_{1,T-3}$	$a_{1,T-2}$	$a_{1,T-1}$	$a_{1,T}$
	2	$a_{2,1}$	$a_{2,2}$...	$a_{2,T-3}$	$a_{2,T-2}$	$a_{2,T-1}$	$a_{2,T}$

	N	$a_{N,1}$	$a_{N,2}$...	$a_{N,T-3}$	$a_{N,T-2}$	$a_{N,T-1}$	$a_{N,T}$

N – число рядов в пучке

T – длина рядов

$$a_{i,j} \in E_k, i = 1, 2, \dots, N, j = 1, 2, \dots, T$$

Пучок k -значных временных рядов $\|a_{i,j}\| \in E_k^{N \times T}$

Задача: поиск «плавно» меняющихся закономерностей

Постоянная закономерность

Закономерность $R = (p, \omega, f)$

$$R = \left\{ \begin{array}{l} 1. p - \text{номер целевого ряда } p \in \{1, 2, \dots, N\} \\ 2. \omega - \text{маска } (||\omega|| - \text{число единиц, мощность}) \\ 3. f - \text{частично определённая функция, } \omega \in E_2^{N \times \Delta} \\ \text{зависящая от } ||\omega|| \text{ переменных} \end{array} \right.$$

1. Ряд p . Например, $p=2$

$a_{1,1}$	$a_{1,2}$...	$a_{1,T-1}$	$a_{1,T}$
$a_{2,1}$	$a_{2,2}$...	$a_{2,T-1}$	$a_{2,T}$
...
$a_{N,1}$	$a_{N,2}$...	$a_{N,T-1}$	$a_{N,T}$

2. Маска ω

		$t-\Delta$	$t-3$	$t-2$	$t-1$	t
N ↑	1	0	0	1	0	
	2	1	0	0	0	
	
	
	N	0	0	1	1	
		Δ (максимальный отступ по времени)				

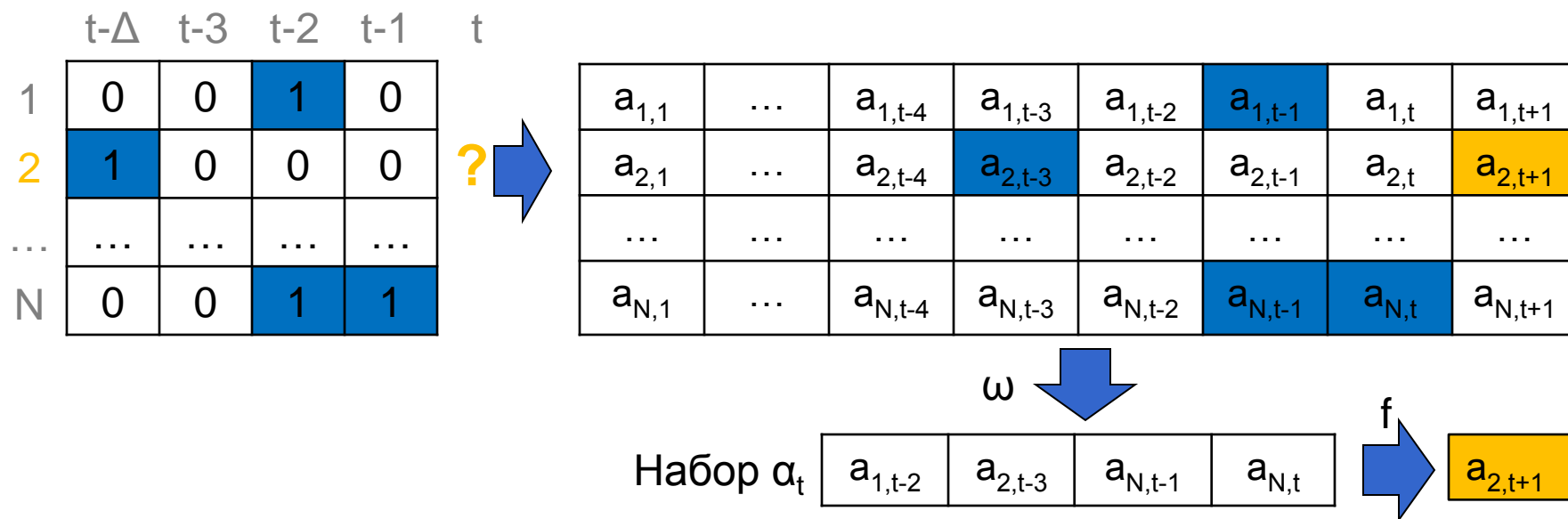
3. Функция f

$$f : E_k^{||\omega||} \rightarrow \{0, 1, \dots, k-1, \lambda\}$$

λ – значение не определено

$$E_k = \{0, 1, \dots, k-1\}$$

Алгоритм поиска постоянных закономерностей



- Предложен алгоритм поиска постоянных закономерностей
- Получены оценки необходимой длины пучка временных рядов для поиска закономерностей, определенных на всех наборах значений аргументов

Оценка необходимой длины пучка временных рядов

Теорема. Пусть все наборы из $E_k^{||\omega||}$ появляются в множестве наборов $\{\alpha_t\}$ с равной вероятностью. Обозначим:

- $M = k^{||\omega||}$ - число всех наборов из $E_k^{||\omega||}$;
- $L = T - \Delta$ - число элементов в множестве $\{\alpha_t\}$ ($M < L$);
- P - вероятность того, что в множестве $\{\alpha_t\}$ присутствуют все наборы из $E_k^{||\omega||}$.

Тогда $P = M! \cdot S(L, M) / M^L$, где $S(L, M)$ - число Стирлинга II-го рода.

Минимальное значение T при $P_0 = 0,95$ и $\Delta = 10$

$k \backslash \omega $	1	2	3	4	5	6
2	16	26	48	100	213	463
3	21	54	177	604	2 063	6 977
4	26	100	463	2 186	10 145	46 241
5	31	162	982	5 886	34 435	197 299

P_0 – вероятность того, что функция f определена на всех наборах

Идея подхода

1. Разбиение исходного пучка на отрезки
2. Поиск закономерностей на каждом из отрезков
3. «Склеивание» «близких» закономерностей различных отрезков → плавно меняющаяся закономерность

Изменяющиеся закономерности

Фиксируем целевой ряд

Время →

Ряд ↓

	1	2	...	T-3	T-2	T-1	T
1	$a_{1,1}$	$a_{1,2}$...	$a_{1,T-3}$	$a_{1,T-2}$	$a_{1,T-1}$	$a_{1,T}$
2	$a_{2,1}$	$a_{2,2}$...	$a_{2,T-3}$	$a_{2,T-2}$	$a_{2,T-1}$	$a_{2,T}$
...
N	$a_{N,1}$	$a_{N,2}$...	$a_{N,T-3}$	$a_{N,T-2}$	$a_{N,T-1}$	$a_{N,T}$

отрезок 1

отрезок 2

отрезок m

Алгоритм поиска постоянных закономерностей

Закономерности:

$$R^1 \in \{R_1^1, R_2^1, \dots, R_{q_1}^1\} \quad R^2 \in \{R_1^2, R_2^2, \dots, R_{q_2}^2\} \quad R^m \in \{R_1^m, R_2^m, \dots, R_{q_m}^m\}$$

$$\tilde{R} = \{R^1, R^2, \dots, R^m\} \text{ - изменяющаяся закономерность} \quad R^i \text{ - шаги } \tilde{R}$$

Меры сходства масок и функций

Закономерности: $R_1 = (p, \omega_1, f), R_2 = (p, \omega_2, g)$

Меры сходства:

- на масках одинаковой мощности $\rho_m(\omega_1, \omega_2)$
- на масках произвольной мощности $\rho_m^\mu(\omega_1, \omega_2)$

Теорема. Отображение ρ_m является метрикой.

«Вспомогательная» мера сходства $\hat{\rho}$ с параметром w

Теорема. Отображение $\hat{\rho}$ является метрикой тогда и только тогда, когда $k \leq 2w+1$.

Следствие. Минимальное w , при котором $\hat{\rho}$ является метрикой равно $(k-1)/2$.

Мера сходства частично-определённых функций ρ_f с параметром w

Теорема. Минимальное w , при котором ρ_f является метрикой равно $(k-1)/2$.

Близкие закономерности

Закономерности: $R_1 = (p, \omega_1, f), R_2 = (p, \omega_2, g)$

Мера сходства закономерностей:

$$\rho(R_1, R_2) = \kappa_m \cdot \rho_m^\mu(\omega_1, \omega_2) + \kappa_f \cdot \rho_f(f, g) \quad 0 \leq \kappa_m \leq 1, 0 \leq \kappa_f \leq 1$$

$$\kappa_m + \kappa_f = 1$$

$$0 \leq \rho_m^\mu \leq 1, 0 \leq \rho_f \leq 1$$

Мера сходства на масках

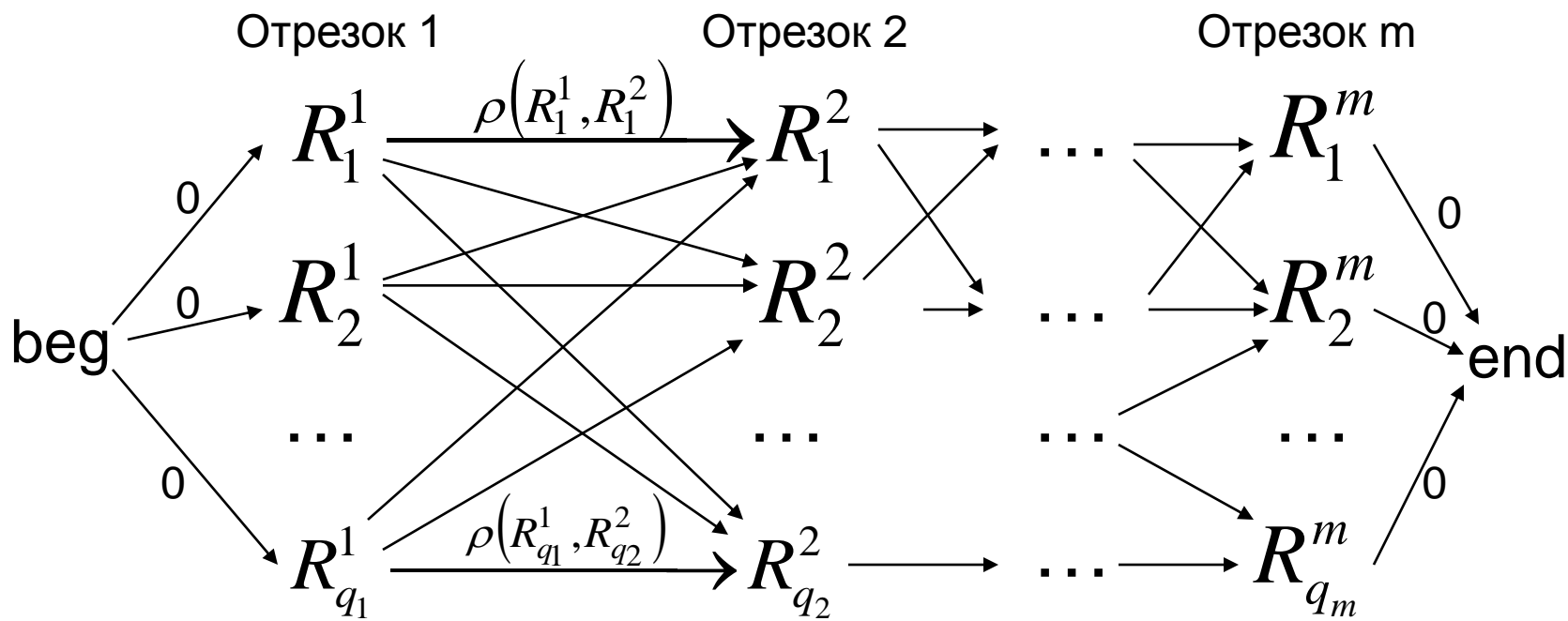
Метрика на частично-определённых функциях

Граф закономерностей

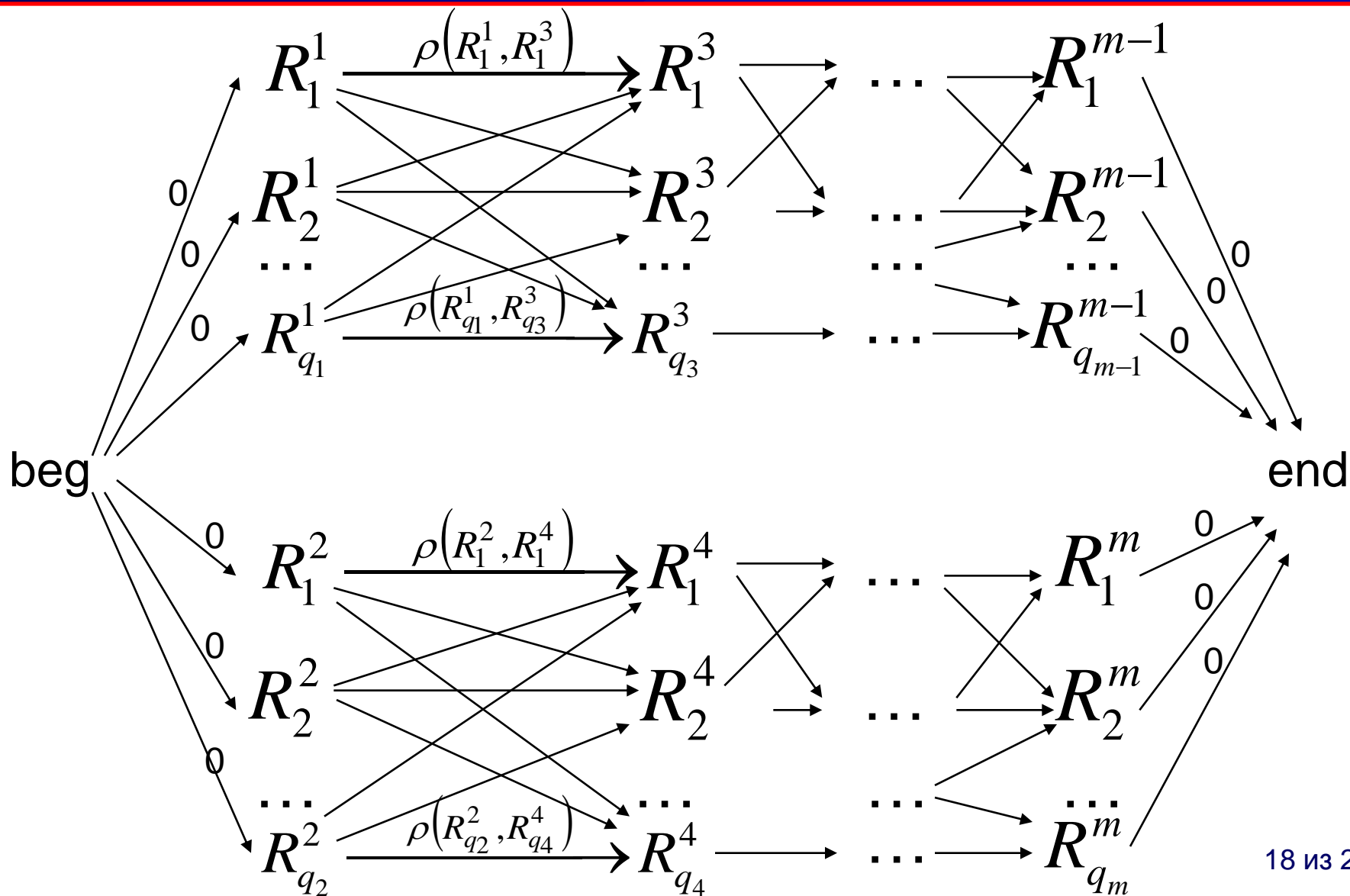
Ориентированный граф со взвешенными вершинами и ребрами:

Вес ребра – мера сходства закономерностей

Весы вершин – функционалы качества постоянных закономерностей

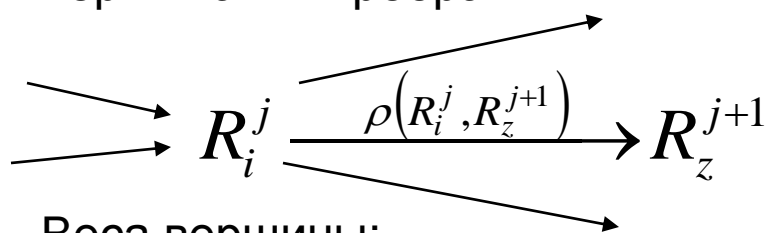


Граф для поиска периодических закономерностей



Функционал качества шага изменяющейся закономерности

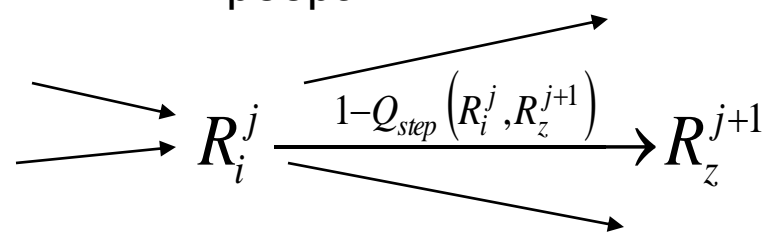
Граф 1 со взвешенными
вершинами и ребрами



Веса вершины:

- Достоверность $Conf(R_i^j)$
- Поддержка $Supp(R_i^j)$

Граф 2 со взвешенными
ребрами



Функционал качества шага изменяющейся закономерности:

$$Q_{step}(R_i^j, R_z^{j+1}) = w_{conf} \cdot Conf(R_i^j) + w_{supp} \cdot Supp(R_i^j) + w_{similarity} \cdot (1 - \rho(R_i^j, R_z^{j+1})) \rightarrow \max$$

Плавно меняющаяся закономерность определяется как кратчайший путь из вершины beg в вершину end на Графе 2

Поиск плавно меняющихся закономерностей

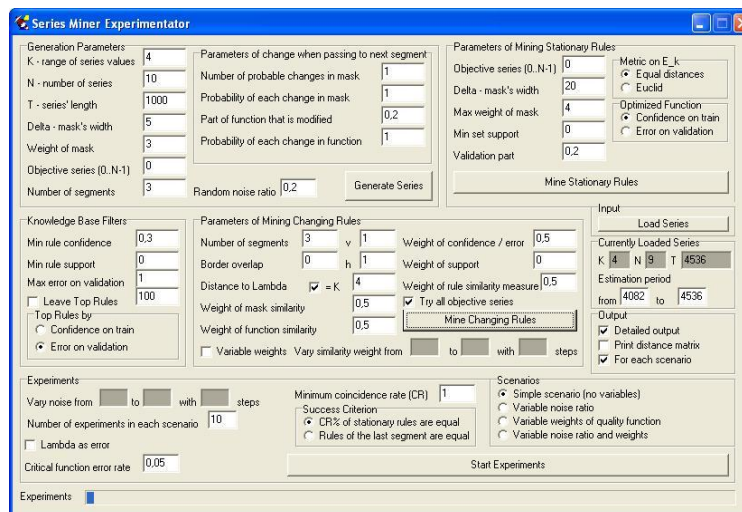
Алгоритм поиска плавно меняющихся закономерностей:

1. Разбить пучок временных рядов на отрезки
2. Произвести поиск постоянных закономерностей на каждом отрезке
3. Построить граф закономерностей и рассчитать меры сходства
4. Найти оптимальный путь на графе закономерностей

Экспериментальный стенд

Экспериментальный стенд позволяет:

- Импортировать и генерировать пучки временных рядов
- Проводить поиск стационарных и изменяющихся закономерностей
- Решать задачи прогнозирования



Эксперименты на модельных данных

Ход экспериментов:

1. Генерировался пучок временных рядов
2. Генерировалась плавно меняющаяся закономерность
3. Целевой ряд заполнялся на основе генерируемой закономерности при заданном уровне шума ε
4. Производился поиск плавно меняющихся закономерностей в сгенерированном пучке временных рядов (при различных весах функционала качества шага изменяющейся закономерности)
5. Рассчитывалась доля успешных экспериментов

Эксперимент называется **успешным**, если найденная закономерность совпадает с генерируемой.

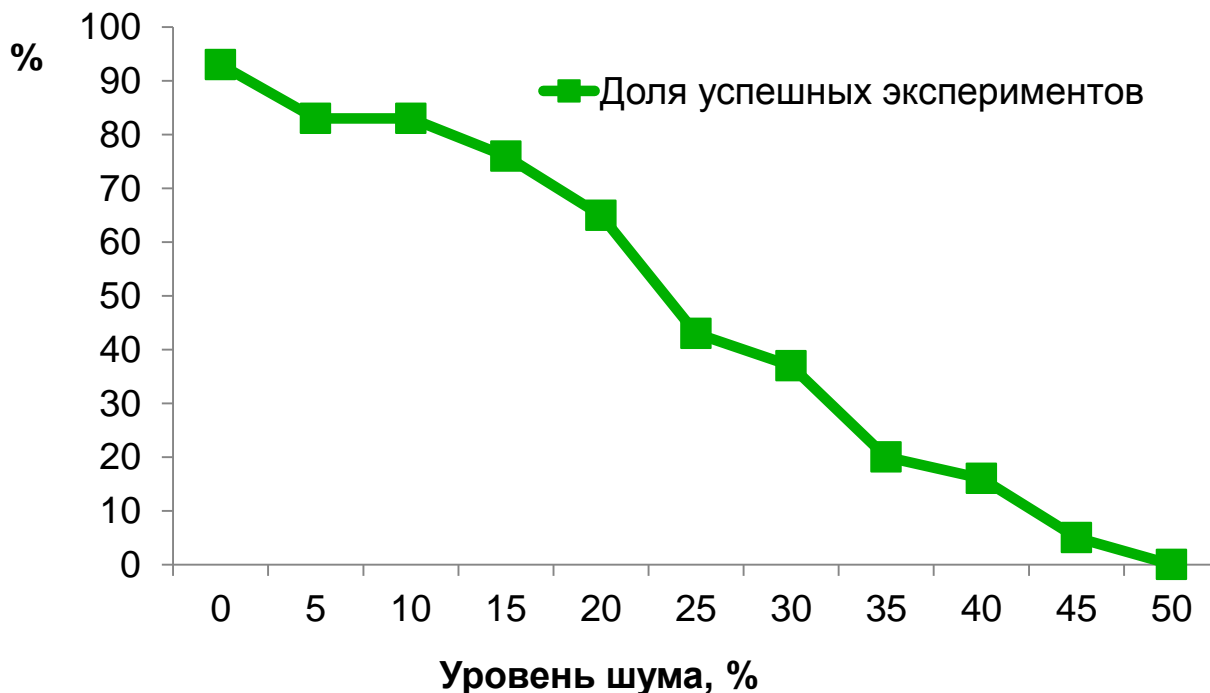
Изменяющиеся закономерности **совпадают**, если совпадают все их соответствующие шаги (постоянные закономерности).

Постоянные закономерности **совпадают**, если полностью совпадают их маски, а функции различаются не более чем на 5% наборов.

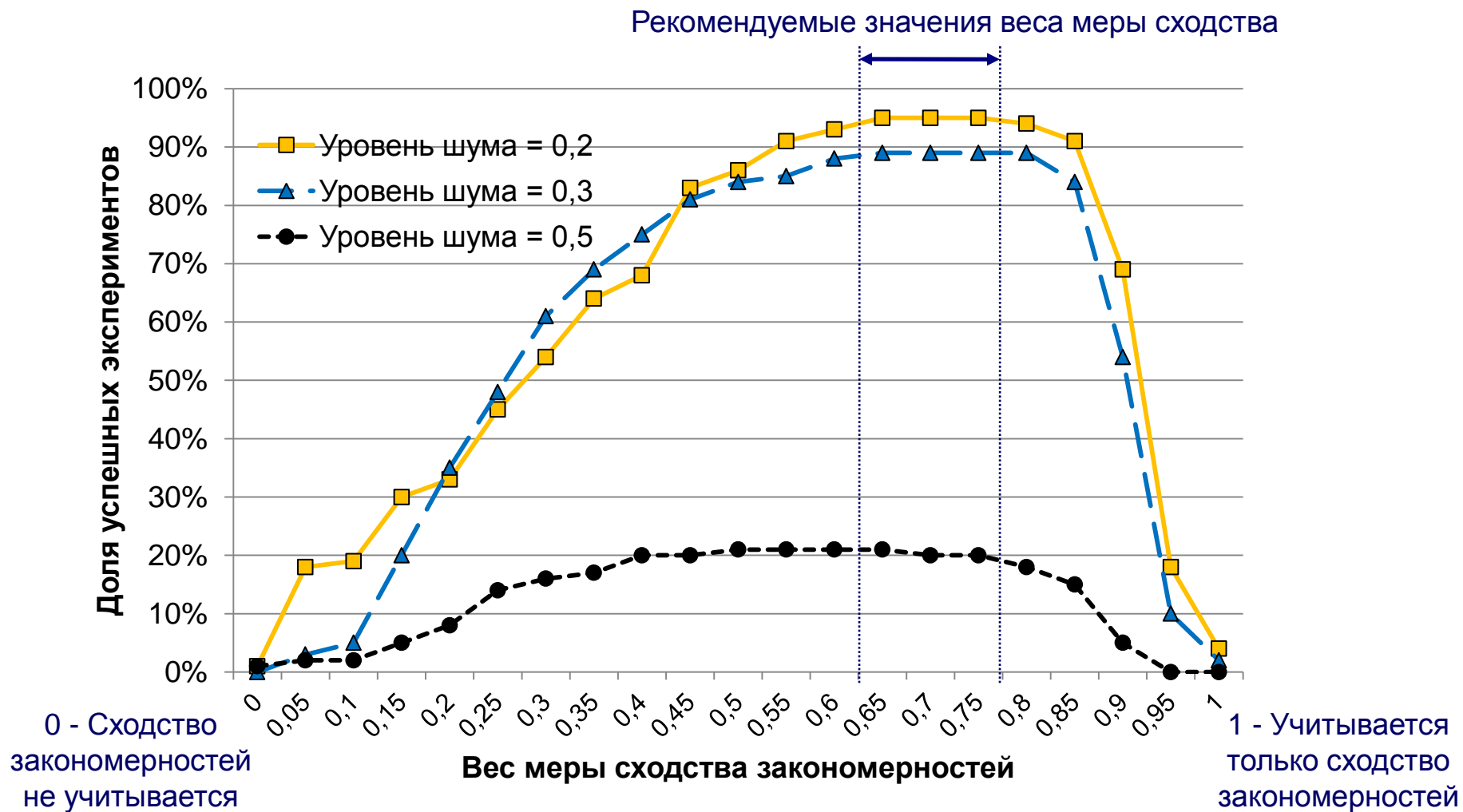
Доля успешных экспериментов при различном уровне шума

Результаты:

- Качество распознавания линейно убывает при увеличении уровня шума
- Алгоритм проводит эффективный интеллектуальный анализ данных даже для зашумленных пучков временных рядов



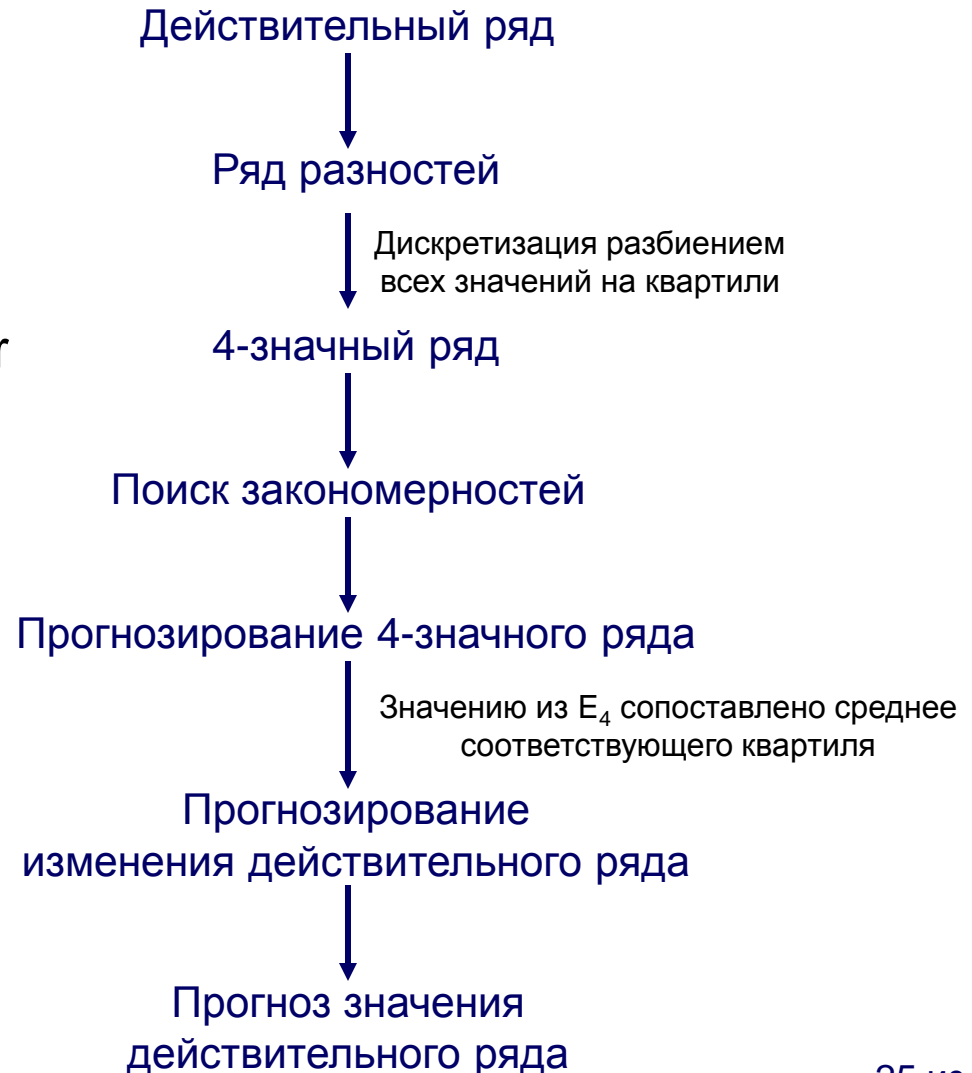
Доля успешных экспериментов при различных весах функционала качества



Анализ реальных пучков временных рядов

Исходные данные:

- Курсы акций компаний, работающих в сфере ИТ (Adobe, BMC, Business Objects, Cognos, Computer Associate, Novell, Oracle, Peoplesoft, Rational).
- Средний почасовой курс акций в долларах за период с 13 мая 2002 по 10 декабря 2004.



Результаты прогнозирования курса акций

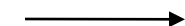
Средний квадрат ошибки

Ряд	Экспон. сгл. ($\alpha=0,3$)	Предложенный алгоритм
Adobe	$7,32 \cdot 10^{-2}$	$6,89 \cdot 10^{-2}$
BMC	$11,15 \cdot 10^{-3}$	$9,72 \cdot 10^{-3}$
Business Objects	$7,19 \cdot 10^{-2}$	$3,74 \cdot 10^{-2}$
Cognos	$3,08 \cdot 10^{-2}$	$2,39 \cdot 10^{-2}$
Computer Associate	$2,87 \cdot 10^{-2}$	$1,91 \cdot 10^{-2}$
Novell	$4,18 \cdot 10^{-2}$	$2,54 \cdot 10^{-2}$
Oracle	$6,77 \cdot 10^{-3}$	$5,45 \cdot 10^{-3}$
Peoplesoft	$2,94 \cdot 10^{-3}$	$1,26 \cdot 10^{-3}$
Rational	$3,43 \cdot 10^{-2}$	$2,06 \cdot 10^{-2}$

Выявленные закономерности

Rational al (t-1)	Comp. Assos. (t-8)	Rational al (t)
0	0	0
0	1	0
0	2	0
0	3	0
1	0	0
1	1	0
1	2	0
1	3	0
2	0	3
2	1	0
2	2	3
2	3	3
3	0	3
3	1	3
3	2	3
3	3	3

Отрезок 1



Компания
Rational
была
куплена
компанией
IBM

Rational al (t-1)	Comp. Assos. (t-6)	Rational al (t)
0	0	0
0	1	0
0	2	0
0	3	0
1	0	1
1	1	1
1	2	0
1	3	1
2	0	2
2	1	1
2	2	2
2	3	2
3	0	3
3	1	3
3	2	3
3	3	2

Отрезок 2



Rational al (t-1)	Comp. Assos. (t-6)	Rational al (t)
0	0	0
0	1	0
0	2	1
0	3	0
1	0	1
1	1	1
1	2	1
1	3	1
2	0	1
2	1	2
2	2	2
2	3	2
3	0	3
3	1	3
3	2	2
3	3	3

Отрезок 3

Значение	0	1	2	3
Символ				
Среднее для Comp. Assos.	-0,16745	-0,02772	0,03438	0,17251
Среднее для Rational	-0,20687	-0,02739	0,03584	0,19421

Публикации

1. *Филипенков Н.В.* Поиск плавно меняющихся закономерностей в пучках временных рядов // Ломоносов-2006: Тез. докл. - М.: Изд-во МГУ, 2006. – С. 56-57.
2. *Филипенков Н.В.* О задачах анализа пучков временных рядов с изменяющимися закономерностями // Искусственный интеллект. – 2006. - №2. - С. 125-129.
3. *Филипенков Н.В.* Об одном методе выявления плавно меняющихся закономерностей в k-значных временных рядах // 49-я конф. МФТИ: Тез. докл. - М.: МФТИ, 2006. – С. 270-271.
4. *Филипенков Н.В.* Об оптимальном выборе закономерностей, составляющих плавно меняющуюся закономерность // ММРО-13: Тез. докл. - М.: МАКС Пресс, 2007. – С. 223-225.
5. *Филипенков Н.В.* Поиск плавно меняющихся ассоциативных правил // 50-я конф. МФТИ: Тез. докл. - М.: МФТИ, 2007. – С. 117-119.
6. *Филипенков Н.В.* Об эволюционирующих алгоритмах классификации и прогнозирования // ИОИ-2008: Тез. докл. - Симферополь, 2008. – С. 228-230.
7. *Филипенков Н.В.* О некоторых аспектах интеллектуального анализа пучков временных рядов // ММРО-14: Тез. докл. - М.: МАКС Пресс, 2009. – С. 204-207.
8. *Филипенков Н.В.* Об одном методе поиска плавно меняющихся закономерностей в пучках временных рядов // **Ж. вычисл. матем. и матем. физ. - 2009. - Т. 49, № 11. С. 2020-2040.**
9. *Filipenkov N. V.* On the mining of slightly changing patterns in multidimensional time series // Proc. Pattern Recognition and Information Processing-2007 - Minsk, 2007. – Vol. 1 – pp. 123-127.
10. *Filipenkov N. V.* Data mining in non-stationary multidimensional time series using a rule similarity measure // Proc. European Conf. on Data Mining-2008 - Amsterdam, 2008. – pp. 92-96.